Szófaji kódok és névelemek együttes osztályozása

Móra György[1], Vincze Veronika[1], Zsibrita János[1]

[1] Szegedi Tudományegyetem,
Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék
6720 Szeged, Árpád tér 2.
{gymora, vinczev, zsibrita}@inf.u-szeged.hu

Kivonat: Jelen munkánkban egy, a szófaji kódok és a névelemek meghatározására szolgáló gépi tanulási modellt mutatunk be. Az általános véletlen mezőkön alapuló módszer segítségével több címkesorozat együttesen tanulható, valamint az osztályozás során a címkesorozatok legjobb kombinációját együttesen keressük. A `magyarlanc` szófaji elemző és az SZTENER névelem-felismerő jellemzőkészletét használva olyan rendszert építettünk, amely a címkék együttes osztályozásának segítségével felülmúlta a kiindulási rendszereket az általunk használt teszthalmazon. A névelem-felismerő F-mértékben mért teljesítménye 87,75-ről 89,87-re, a szófaji címkéző pontossága 97,11%-ról 97,99%-ra nőtt, úgy, hogy a kódok meghatározásának más minőségi tényezői is javultak.

1 Bevezetés

Szintaktikai szempontból a tulajdonnevek főnévként viselkednek: a *Láttad az **Interjú a vámpírral**-t?* mondatban a film címe ugyanúgy ragozható, mint bármely más magyar főnév (vö. *Láttad a filmet?*). Emiatt a tulajdonneveket gyakran a főnevek egyik alosztályának tekintik: bizonyos morfológiai kódrendszerek külön tulajdonnévi kódot tulajdonítanak nekik (például az MSD-kódrendszerben Np-s*, a PENN Treebankben pedig NNP az egyes számú tulajdonnevek kódja).

Azonban valójában nemcsak főnevek, hanem bármelyik szófajhoz tartozó elemek is lehetnek tulajdonnevek (vagy azok részei), például *Tesz-Vesz Kft*. A fenti kódrendszerek használatával a *Tesz-Vesz*-t is tulajdonnévnek kellene kódolni, ami azonban a kódok megsokszorozódásával jár, hiszen voltaképpen bármely szónak lehet tulajdonnévi kódja is. Ez egyrészt megnöveli a szófaji egyértelműsítés költségeit (sokkal több szó válik morfológiailag többértelművé), továbbá megkívánja azt is, hogy a morfológiai elemzőbe beépüljön egy tulajdonnév-felismerő rendszer. Úgy véljük azonban, hogy a tulajdonnév-felismerés nem a morfológiai elemző feladata, így az általunk alkalmazott megoldásban a két feladatot párhuzamosan hajtjuk végre. Megközelítésünkben a tulajdonnévi jelölés tehát nem a morfológiai kód része, hanem külön tulajdonnévi címkékkel látjuk el a tulajdonnév-felismerő által NE-nek ítélt elemeket, függetlenül attól, hogy milyen szófajú az adott elem.

Munkánkban megmutatjuk, hogy a szófaji címkézés és a névelem-felismerés teljesítménye kölcsönösen javítható a tanulás során a másik feladat által szolgáltatott jelö-

lésekkel. Hogy ez lehetővé váljon, olyan gépi tanuló megközelítést alkalmaztunk, amelynek segítségével a két probléma együtt, egy gépi tanulási faladatként kezelhető. Az általunk fejlesztett rendszer hatékonyan alkalmazható magyar nyelvű szövegek egyidejű szófaji címkézésére és a bennük található névelemek felismerésére, és a használt tanító és kiértékelő halmazokat figyelembe véve teljesítményében felülmúlja az eddigi különálló statisztikai alapú szófaji címkézőket, valamint névelem-felismerő rendszereket. A módszer könnyen adaptálható más nyelvekre is, amennyiben rendelkezésre áll az adott nyelven morfológiai elemző és megfelelő annotált szövegkorpusz, mivel nem alkalmaz nyelvspecifikus jellemzőket.

2 Morfológia és tulajdonnevek

A tulajdonnevek nyílt szóosztályt alkotnak, azaz nem alkotnak véges elemű halmazt, számuk állandóan bővül a nyelvben. Ez maga után vonja, hogy nem is sorolhatók fel maradéktalanul egy szótárban sem. A nyelvfeldolgozás számára azonban kiemelkedően fontos a tulajdonnevek megfelelő kezelése, így például a morfológiai elemzőkbe nagyméretű tulajdonnévszótárak épülnek be azok elemzésének megkönnyítésére. Azonban a fenti okok miatt egy morfológiai elemző sem ismerhet fel minden szóalakot, így az ismeretlen szavak (melyek nagy része tulajdonnév vagy annak származéka) kezelésére különféle, úgynevezett guessing módszereket érdemes kidolgozni [20].

A tulajdonneveket a nyelvészeti szakirodalom többnyire merev jelölőnek tekinti, mely konstans módon ugyanazt az egyedet azonosítja [7]. A fenti definícióban a „merevség” arra vonatkozik, hogy nem változik a jelölő és jelölt közti kapcsolat, azonban elgondolásunk szerint a „merevség” fogalma a tulajdonnevek morfológiájában is értelmezhető. A tulajdonnevek ugyan ragozhatók, sőt alkalmanként képzők is csatlakozhatnak hozzájuk (*New York – New York-i*), azonban a lemmájuk változatlan formában fordul elő a toldalék előtt (*Fodor – fodoros*). (A kisbetű-nagybetű változásoktól most eltekintünk.) Ez különösen akkor nyilvánvaló, amikor egy morfológiailag sajátos viselkedésű főnév fordul elő tulajdonnévi használatban. Vegyük az alábbi példákat.

Fodort Kovács, míg Bokort Szabó váltotta az elnöki székben.
Panni átugrotta a bokrot, és egy kiálló ág elszakította a szoknyája alján levő fodrot.

A *fodor* és *bokor* hangkivető főnevek, vagyis bizonyos toldalékok előtt kiesik a lemma utolsó magánhangzója. Ez a jelenség azonban nem figyelhető meg akkor, amikor személynévként használatos a két szó. E tulajdonság kihasználható a névelem-felismerésben: a morfológiai elemző a *fodrot* és *bokrot* alakokat várná *fodr+ot* és *bokr+ot* morfémákkal, ám a fenti szóalakokat csak a guesser segítségével lehet elemezni a beépített toldaléklista segítségével *fodor+t*, illetve *bokor+t* morfémákra való felbontással. Amennyiben az így kapott lemma megtalálható a morfológiai adatbázisban, viszont eltérést tapasztalunk az ott található és a guesser által adott elemzés között (vagyis jelen esetben a *fodor* és *bokor* tárgyesetű alakja nem *fodrot* és *bokrot*, hanem *fodort* és *bokort*), valószínűsíthetjük, hogy tulajdonnévről van szó.

Bizonyos tulajdonnévtípusok – műcímek, intézménynevek (különösen ha többtagúak) – gyakran tartalmaznak már eleve ragozott alakokat, például *Interjú a vámpírral*, *Bolyai Farkas Alapítvány a Magyarul Tanuló Tehetségekért*. Azonban ezek is ragozhatók:

Megnéztem az Interjú a vámpírral-t.
Köszönetet mondott a Bolyai Farkas Alapítvány a Magyarul Tanuló Tehetségekért-nek.

A helyesírási szabályok szerint ilyenkor kötőjellel kell kapcsolni az újabb toldalékot a tulajdonnévhez. Utóbbi sajátosság is kihasználható a névelem-felismerésben: a kötőjelet tartalmazó szóalakot a guesser segítségével elemezzük, majd az így kapott lemmát ismét elemezzük. Amennyiben a szóalak a második elemzés során is toldalékoltnak bizonyul, ismét valószínűsíthető, hogy tulajdonnévvel találkoztunk.

A gyakorlatban sokszor előfordul, hogy a toldalék nem kötőjellel kapcsolódik a tulajdonnévhez (akár a helyesírási szabályok ellenében). Ezekben az esetekben is a guesser nyújthat segítséget: a lehetséges végződéseket le kell vágni a szó végéről, majd a maradékot lemmaként visszaadni, és a toldaléknak megfelelő főnévi elemzést társítani a szóhoz (pl. *Agrobankhoz – Agrobank* illativusi esetű főnév).

A morfológiai elemző oldaláról nézve a vele párhuzamosan zajló tulajdonnévfelismerés abban segíthet, hogy a NER-rendszer által tulajdonnévnek minősített elemeket nem feltétlenül próbálja meg hagyományos módon elemezni, hanem egyből a beépített guessert hívja segítségül, ezzel gyorsítva a folyamatot.

3 Együttes címkézési módszerek

Hagyományosan a különböző szekvenciajelölési feladatokat (szófaji címkék, felszíni elemzés, névelemek) külön-külön gépi tanulási feladatként definiálják, és a szövegek feldolgozása során az elemzőket egymás után futtatják. Így azonban az egyes alrendszerek hibái összeadódnak, valamint csak a feldolgozási láncban hátrébb álló komponenseknek van lehetősége felhasználni az előtte állók címkéit jellemzőként.

3.1 A címketerek kombinálása

Több jelölési lépés egyesíthető a címkék kombinálásával is, de így kezelhetetlen mértékben megnőhet a címketér, illetve előfordulhat, hogy bizonyos címkekombinációk csak kevésszer fordulnak elő a tanuló adatok között, így felismerésük bizonytalan lesz. A feladatok ilyen jellegű kombinálásánál a közös jellemzőkészlet is problémát jelenthet, mert előfordulhat, hogy a különböző címkézési feladatok eltérő jellemzőkészlet mellett adnak optimális eredményt.

3.2 Gráfalapú valószínűségi modellek

Kísérleteinkben a szövegek párhuzamos címkézésére a MALLET GRMM [9][15] és a FactorIE [11] csomagban található általános feltételes véletlen mezők módszerét alkalmaztuk. A módszerek lehetővé teszik a hagyományos lineáris láncolású véletlen mezők módszeréhez képest, hogy tetszőleges valószínűségi függőségeket ábrázoló modelleket alkalmazzunk, így egy token akár egynél több címkével is rendelkezhet. A címkék közötti feltételes valószínűségi kapcsolatok modellezésével a névelem-felismerés és a szófaji címkézés egymástól független jellemzőkészlet segítségével valósítható meg, de olyan módon, hogy a szófajcímkék és a névelemcímkék együttes legjobb eloszlását tanuljuk, majd keressük a jelölés során. Természetesen a módszer kiterjeszthető más feladatokra, vagy akár kettőnél több egyidejű címkesorozat meghatározására is.

3.3 Előzetes vizsgálatok

Angol nyelvű szövegeken végzett kísérletek [10] azt mutatták, hogy a szófaji kódok és a felszíni elemzés címkéinek együttes gépi tanulásával jobb eredményt lehet elérni, mint ha ezeket a feladatokat külön tanított modellekkel egymás után szekvenciálisan végeznék el. Az általunk végzett ilyen irányú kísérletek azt mutatták, hogy a szófaji kódok meghatározásának pontossága 62,45%-ról 72,89%-ra, a felszíni elemzés pontossága pedig 83,95%-ról 85,76%-ra nőtt azonos jellemzőkészlet használata mellett, abban az esetben, ha a címkesorozatokat független osztályozása helyett azokat együttes osztályozással határozzuk meg. A két címkesorozat az osztályozás során így dinamikus jellemzőként hathat egymásra, kölcsönösen javítva a címkék meghatározásának pontosságát. A mérésekhez a CoNLL-2000 Shared Task tanító és kiértékelő halmazának ezer-ezer tokenes mintáját használtuk.

A CoNLL-2003 Shared Task [18] nyelvfüggetlen névelem-felismerési feladatán végzett kísérletek azt mutatták, hogy minimális jellemzőkészletet használva, mind a szófaji kódok címkézése, mind a névelemek felismerése javítható az együttes címkézés használatával. A verseny spanyol szövegeket tartalmazó részkorpuszából származó mintán elvégzett vizsgálatok azt mutatták, hogy míg a szófaji kódok címkézésének pontosságát csak mérsékelten 88,6%-ról 88,7%-ra, addig a névelem-felismerés F-mértékét jelentős mértékben, 39,5-ről 42,2-re növelte az együttes címkézés.

4 Névelem-felismerés

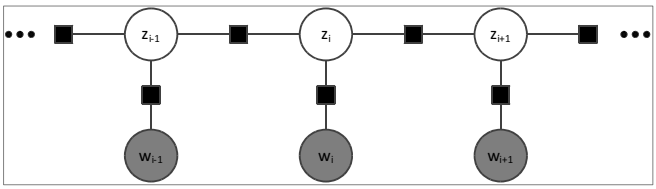
A névelem-felismerés alapvető fontosságú az információkinyerő rendszerek működése szempontjából. A felismert és különböző típusokba sorolt névelemek nem csak önmagukban érdekesek, de sok rendszerben a névelemek jelentik azokat az alapegységeket, amelyekből események épülnek fel, illetve amelyek között relációkat azonosítanak. A névelemek azonosításánál általában sokkal nagyobb kihívást jelent azok megfelelő osztályba sorolása. Az osztályozás általában környezeti jellemzők alapján lehetséges.

4.1 Kapcsolódó munkák

A névelemek felismerésének két alapvető módját különböztethetjük meg. A tokenalapú rendszerek szavankénti osztályozással döntik el, hogy az adott token része-e vagy sem egy névelemnek. Az osztályozó rendszerint szupportvektorgép [8], vagy maximum entrópia osztályozó [1][5]. Gyakran több akár különböző típusú tanulót is kombinálnak [13]. A névelem-felismerők másik, elterjedtebb csoportja a szekvenciatanulást alkalmazó módszerek. A Markov-mezőket [14] egyre inkább a feltételes véletlen mezők váltják fel a szekvenciajelölő rendszerekben. A CoNLL-2002 és a CoNLL-2003 névelem-felismerési feladatainak eredményei azt mutatták hogy a tokenenkénti osztályozást végző rendszereket többnyire felülmúlják a több token feletti címkeeloszlást tanuló megközelítések a névelem-felismerési feladatokban. [17][18]

Az általunk fejlesztett névelem-felismerő módszer az SZTENER [3] nyelvfüggetlen névelem-felismerő rendszer magyar nyelvre adaptált változatából indul ki. A szoftver a feltételes véletlen mezők módszerének MALLET [9] programcsomagban található verzióján alapszik. Elsőrendű láncolást alkalmaz, a jellemzők között ortografikus, szófrekvencia alapú, valamint szótár jellemzők találhatóak. A tanító és teszthalmaz mondataiból és szavaiból ennek a rendszernek a jellemzőkinyerő modulja segítségével készítettünk a gépi tanuló algoritmusok számára feldolgozható jellemzővektorokat.

4.2 A névelemfelismerő rendszer modellje



1. ábra: A névelemek felismeréséhez használt elsőrendű modell. A fehér körök a címkék rejtett változóit, a szürkék a jellemzők megfigyelhető változóit, a fekete négyzetek a változók közötti faktorokat jelölik.

A névelem-felismerő architektúráját megtartva a FactorIE feltételes valószínűségi programozási környezetben az [11] ábrán látható elsőrendű feltételes valószínűségi modellt definiáltunk. A modell a szó jellemzői $(w_0,w_1,...,w_n)$ és címkéi $(z_0,z_1,...,z_n)$, valamint az egymást követő címkék között definiál faktorokat. Az egyetlen különbség az eredeti és az általunk fejlesztett rendszer között, hogy a feltételes valószínűségek pontos kiszámítása helyett közelítő módszereket alkalmaztunk, ugyanis az együttes címkézési feladat során előálló bonyolult modell kiszámítása csak közelítő módszerekkel kivitelezhető elfogadható időn belül.

5 Szófaji kódok meghatározása

A szófaji kódok fontos szerepet töltenek be a szöveg további nyelvészeti elemzése során, illetve sok megközelítés közvetlenül jellemzőként is használja. A kódok hozzárendelése tokenalapon történik. Jelen munkában az MSD-kódrendszer egy egyszerűsített, gépi tanulási módszerekkel könnyebben kezelhető változatát használjuk.

5.1 Kapcsolódó munkák

Korábban több szófaji címkéző rendszer is készült a magyar nyelvre, mint például a szabály alapú RGLearn, illetve más, rejtett Markov-modellekre épülő statisztikai módszereket alkalmazó algoritmusok [4][6][12]. A szófaji címkézési feladat szerves része – különösen erősen agglutináló nyelvek esetében, mint például a magyar – a szavak morfológiai elemzése. A korábban említett magyar szófaji egyértelműsítők a HuMOR[1], illetve MetaMorpho[2] rendszereket, valamint a NooJ magyarra átültetett verzióját[3] alkalmazták.

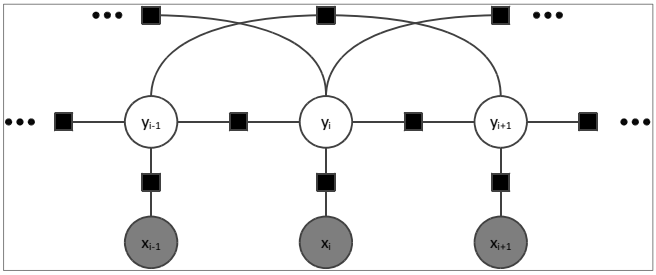
A szófaji címkéző jellemzőkészlete és felépítése a `magyarlanc` nevű [20], a Stanford POSTagger [19] módosításával létrehozott szófaji címkézőn alapszik, amely körkörös függőségű véletlen mezőket alkalmazó maximum entrópia osztályozót használ. A magyar nyelvre kifejlesztett jellemzőkészlet az 1-3 hosszú karakterprefixeket és suffixeket, a szavakat és azok szómintáját tartalmazza. Ezen kívül környezeti jellemzőként a szó előtte és utána álló szavakkal alkotott bigramjait, valamint a szavak és a környezetében található szavak szófaji címkéinek kombinációit használja. A szófaji kódok, illetve azok bi- és trigramjai a címkézés során dinamikusan állnak elő, a rendszer a lehetséges kombinációkat elemezve dönt a címkékről, így a módszer a tokenosztályozás és a szekvenciaosztályozási módszerek jegyeit is magán hordozza. Az adott szóhoz rendelhető szófaji kódokat a morfológiai elemző által megadott lehetséges kódok halmazából veszi a címkéző, ezzel is csökkentve a keresési teret [4].

5.2 A szófaji címkéző modellje

Mivel a szófaji címkéző ciklikus helyi függőségeket tartalmazó maximum entrópia osztályozót használó modellje egy az egyben nem ültethető át a FactorIE feltételes valószínűségi programozási környezetbe, a *2. ábrán* látható, az eredeti módszer ötleteit felhasználó másodrendű véletlen mezős modellt definiáltunk. A modell a névelem-felismerő szerkezetéhez hasonló, de a szó jellemzői $(x_0,x_1,...,x_n)$ és címkéi $(y_0,y_1,...,y_n)$, valamint az egymást követő címkék közötti faktorokon kívül a nem közvetlenül egymást követő címkék között is létrehoz feltételes kapcsolatokat. Ez azért

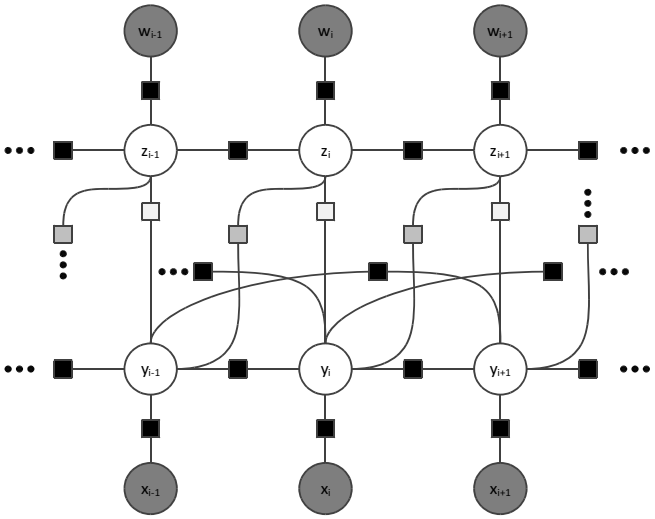
1 http://www.morphologic.hu/Morfologiai-elemzes.html
2 http://www.morphologic.hu/MetaMorpho-technologia/menuazonosito-256.html
3 http://corpus.nytud.hu/nooj/

fontos, mert a szófaji kódok erősen függenek nem csak az őket közvetlenül megelőző, hanem az azt megelőző címkétől is.



2. ábra: A szófaji címkéző által alkalmazott másodrendű modell.

A szavak felszíni jellemzői mellett a morfológiai elemző által megadott lehetséges szófaji kódok is külön vektorváltozóba kerültek. Az eredeti `magyarlanc`tól való eltérés, hogy a keresés nem korlátozódik csak ezekre a címkékre, emiatt számos esetben olyan címkéket is helyesen meghatározott, amiket a morfológiai elemző – hibásan – nem ajánlott fel.



3. ábra: A két különálló valószínűségi modell egyesítése. A világos és sötétszürke színnel jelölt faktorok a két címkesorozat közötti összefüggések leírására szolgálnak.

6 Névelemek és szófaji kódok együttes címkézése

A szófaji címkézés és a névelem-felismerés valószínűségi modelljeit a *3. ábrán* látható modellben egyesítettük. A két címkesorozat elemei között, valamint a névelem címkéjének változója és a megelőző szó szófaji kódjának változója között új faktorokat alkalmaztunk a modellek összekapcsolására. Ezen faktorok paraméterei lesznek azok, amelyek a tanulás után leírják a két címkesorozat közötti összefüggéseket.

7 Eredmények

Módszerünket a Szeged Korpusz üzleti híreket tartalmazó alkorpuszán értékeltük ki, melyben be vannak jelölve az etalon tulajdonnevek [2][16]. Az eredeti MSD-annotációban a tulajdonnevek Np-s* kóddal rendelkeztek, továbbá a többtagú tulajdonnevek össze voltak vonva. A kiértékelést megelőzően szétdaraboltuk a többtagú tulajdonneveket, és tagjaikat újraannotáltuk, a főnevek esetében pedig nem tettünk különbséget a köznévi és tulajdonnévi használat között (azaz a *köznév* és *tulajdonnév* kódokat felváltotta a *főnév* kód). Így tehát a *Magyar Nemzeti Bank* új kódja *A A N* lett. A magyar nyelven végzett kísérleteink azt mutatják, hogy – az angolhoz hasonlóan – eredményeink meghaladják a szekvenciálisan tanított modellek hatékonyságát.

A tanításhoz és a kiértékeléshez a rendelkezésre álló több mint 221 ezer tokent és 9400 mondatot tartalmazó korpuszt két részre osztottuk a mondatok véletlenszerű halmazba sorolásával. A tanító halmazba így a mondatok megközelítőleg 60%-a került, a maradékot kiértékelésre használtuk.

7.1 A névelem-felismerés kiértékelése

A jelen munkában szereplő névelem-felismerésre vonatkozó eredmények mind frázisalapú kiértékelésből származnak. Ez azt jelenti, hogy többszavas névelemek esetén csak az a jelölés számított helyesnek, ahol a névelem minden szava helyesen volt jelölve, és további szavak nem kerültek jelölésre. Az összehasonlíthatóság érdekében az összes rendszert ugyanazokon a halmazokon tanítottuk és értékeltük ki, azonos metrikákat alkalmazva. Ezt a frázisalapú F-mértéket alkalmazták a CoNLL-2003 névelem-felismerési feladat kiértékelése során is, az itt közölt eredmények azonos módszerrel lettek megállapítva.

A kiindulási rendszer teljesítménye mellett az általunk fejlesztett rendszerek eredményeit a tanuló algoritmus 2 és 5 iterációig tartó futtatása mellett is megadjuk mind a szófaji címkézéstől függetlenül tanított névelem-felismerő, mind az együttesen tanított és osztályozott névelem-felismerés esetében.

1. táblázat: Névelem-felismerés eredményei.

It.	Rendszer	Precízió	Fedés	$F_{\beta=1}$
	SZTENER névelem-felismerő	86,81	88,71	87,75
2	Független osztályozás	86,81	81,11	83,86
	Együttes osztályozás	88,57	89,27	88,93
5	Független osztályozás	84,73	81,60	83,13
	Együttes osztályozás	**89,71**	**90,04**	**89,87**

Az *1. táblázatban* található eredmények megerősítik, hogy a névelemek szófaji kódokkal való együttes osztályozása azonos jellemzőtér esetében jelentősen javítja a címkézés teljesítményét a függetlenül tanított modellhez képest. A független modell a kiindulási rendszernél is gyengébb teljesítményét 83,86-ról 88,93-ra növeli. A jellemzőtér ábrázolásának gyengeségét sejteti, hogy az eredetileg is gyengébb eredményt csak csökkenti a tanuló iterációs számának növelése, vélhetően túltanulja a jellemzőket. Ezt az információhiányt kompenzálhatja az együttes tanuláskor a szófaji kódok jelenléte.

7.2 A szófaji címkézés kiértékelése

A szófaji címkézést a csökkentett MSD szófaji kódok alapján tanítottuk és predikáltuk [20]. Ez az MSD-kódoknak egy szűkített készlete (42 kód), ahol csak azok a szófaji kódok vannak megkülönböztetve, ahol a szóalakból nem dönthető el egyértelműen a szó eredeti MSD-kódja. Erre a címketér csökkentése miatt van szükség, mert az eredeti több száz címkét tartalmazó kódrendszer gépi tanuló módszerekkel kezelhetetlen lett volna.

A csökkentett MSD-kódokat tovább redukálva csak a szófajt jelölő első karaktert megtarva is elvégeztük a szófaji címkézők kiértékelését, így láthatóvá vált, hogy a csökkentett MSD-kódokon szinte azonos eredményt elért rendszerek által hibásan jelölt MSD-kódok mennyire térnek el egymástól, azaz mennyire súlyos hibákat vét a két címkéző.

A szófaji címkézést a névelem-felismeréshez hasonlóan a kiindulási rendszerhez hasonlítottuk, és megmértük a csak szófaji címkézést végrehajtó modell és az együttes osztályozás közötti különbségeket is. A rendszerünket ebben az esetben is kettő, illetve öt iterációig tanítottuk.

A névelem-felismeréstől eltérően nem F-mértéket, hanem pontosságot alkalmaztunk a rendszerek teljesítményének elsődleges méréséhez. A pontosság mellett az egyes MSD/szófaji osztályokon elért F-mértékek átlagát (makroátlag, *1. képlet*) is megadtuk a rendszerekhez. Míg a pontosság a szöveg szavainak átlagos osztályozási pontosságát írja le, a makroátlag azt mutatja meg, hogy a ritkán előforduló címkék osztályait mennyire jól ismeri a rendszer. Ha ugyanis csak a gyakori szófajcímkéket osztályozza helyesen, akkor az osztályonkénti F-mértékek átlaga alacsony lesz a sok kis elemszámú, rosszul címkézett szófaji osztály miatt.

2. táblázat: Szófaji címkézés eredményei.

It.	Rendszer	Redukált MSD-kód Pontosság	$F_{\beta=1\ macro}$	Csak szófaj Pontosság	$F_{\beta=1\ macro}$
	magyarlanc	97,11	67,81	97,98	85,18
2	Független oszt.	97,75	71,03	98,60	84,12
	Együttes oszt.	97,78	72,48	98,68	86,32
5	Független oszt.	**98,00**	71,33	98,78	86,44
	Együttes oszt.	97,99	**73,32**	**98,81**	**88,77**

$$F_{\beta=1\ macro} = \frac{\sum F_{\beta=1}(c_i)}{|C|}, \forall c_i \in C \qquad (1)$$

A szófaji egyértelműsítés terén azt tapasztaltuk, hogy eredményeink javulása elsősorban a nagybetűvel kezdődő alakok helyes elemzésének köszönhető. Ez nem meglepő, hiszen a magyarban általában a tulajdonnevek és a mondatkezdő szavak kezdődnek nagybetűvel. A tulajdonnevek és szófaji kódok együttes jelölésével a mondatkezdő tulajdonneveket könnyebb volt azonosítani, így a „maradék" mondatkezdő elemek szófaját is nagyobb hatékonysággal lehetett megállapítani: például a *Szerinte* mondatkezdő elem főnévi kódot kapott a szekvenciális jelölésben, azonban az együttes jelölés során már a helyes határozószói kódot kapta.

Kiemelkedő javulást figyelhettünk meg a rövidítések esetében is. Noha ez a szóosztály kevés elemet tartalmaz, felismerésük 17,86%-kal javult, ami főleg a tulajdonnév részét képező *Jr*. és *Dr*. elő-, illetve utótagoknak pontosabb azonosításának volt köszönhető. Az indulatszavak kategóriájába lettek sorolva olyan tulajdonnevek is, amelyeket a morfológiai elemző – helytelenül – olyan összetételként értelmezett, amelynek utótagja indulatszó, például *Palotainé*. Ezek tulajdonnévként való felismerése javított a rendszer teljesítményén.

Összességében azt figyelhettük meg, hogy a rendszer különösen a ritkán előforduló szófajok felismerésében volt képes javulni, míg a nagyobb szóosztályok esetében minimális különbségeket vehettünk észre. Utóbbiak felismerési pontossága azonban már a szekvenciális modell esetében is kiemelkedő volt (97% feletti), így a tulajdonnevek hozzáadott értéke nem befolyásolta érdemben az eredményeket.

Az elhanyagolható pontosságbeli eltérés ellenére a jelölés minősége javult az együttes osztályozástól. A *2. táblázatban* található makroátlagok azt mutatják, hogy közel azonos pontosság mellett az együttesen tanított rendszer a kis elemszámú szófaji kódok osztályozásában jobb, ezzel összességében kiegyensúlyozottabb teljesítményt nyújt. A hibaelemzéshez alkalmazott, csak a szófajt figyelembe vevő kiértékelés pedig azt mutatja, hogy az együttesen tanított rendszer hibás címkézéskor több esetben rendel olyan szófaji kódot a szavakhoz, amelyek szófaja megegyezik a helyes szófajjal, azaz az elkövetett hibáinak kisebb hányada súlyos tévesztés, mint a függetlenül tanított szófaji kódcímkézőnek.

8 Konklúzió

Cikkünkben a szófaji kódok és a névelemek együttes címkézéséhez használható rendszert mutattunk be. Megmutattuk, hogy a hagyományos, szeparáltan tanuló módszerekhez képest mindkét címkézési feladat teljesítménye nőtt. Bár a szófaji címkézés esetében a változás nem olyan jelentős, de javultak az egyéb minőségi tulajdonságai.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER és BELAMI kódnevű projektek keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Borthwick, A.: Maximum Entropy Approach to Named Entity Recognition. PhD thesis, New York University (1999)
2. Csendes D., Hatvani Cs., Alexin Z., Csirik J., Gyimóthy T., Prószéky G., Váradi T.: Kézzel annotált magyar nyelvi korpusz : a Szeged Korpusz. In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003). Szeged (2003) 238–247
3. Farkas R., Szarvas Gy.: Nyelvfüggetlen tulajdonnév-felismerő rendszer, és alkalmazása különböző domainekre. In: Alexin Z., Csendes D. (szerk.): IV. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2006) 22–31
4. Halácsy P., Kornai A., Oravecz Cs.: HunPos — an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (2007)
5. Chieu, H. L., Ng, H.T.: Named Entity Recognition with a Maximum Entropy Approach. In: Proceedings of CoNLL-2003 (2003)
6. Kuba A., Bakota T., Hócza A., Oravecz Cs.: A magyar nyelv néhány szófaji elemzőjének összevetése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 16–22
7. Kripke, S.: Naming and necessity. Blackwell, Oxford (1980)
8. Mayfield, J., McNamee, P., Piatko, C.: Named Entity Recognition using Hundreds of Thousands of Features. In: Proceedings of CoNLL-2003 (2003).
9. McCallum, A,. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. (2002).
10. McCallum, A., Rohanimanesh, K., Sutton,C.: Dynamic Conditional Random Fields for Jointly Labeling Multiple Sequences. In: NIPS Workshop on Syntax, Semantics and Statistics (2003)
11. McCallum, A., Schultz, K., Singh, S.: FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs. In: Advances on Neural Information Processing Systems (NIPS) (2009)
12. Novák A., Nagy V., Oravecz Cs.: Magyar ismeretlenszó-elemző program fejlesztése. In: Alexin Z., Csendes D. (szerk.): I. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2003) 45–54

13. Radu, F., Ittycheriah, A., Jing, H., Zhang, T.: Named Entity Recognition through Classifier Combination. In: Proceedings of CoNLL-2003 (2003)
14. Miller, S., Crystal, M., Fox, H., Ramshaw, L., Schawartz, R., Stone, R., Weischedel, R. and the Annotation Group: BBN: Description of the SIFT System as Used for MUC-7. In: MUC-7. Fairfax, Virginia (1998)
15. Sutton, C.: GRMM: GRaphical Models in Mallet..http://mallet.cs.umass.edu/grmm/.
16. Szarvas, Gy., Farkas, R., Felföldi, L., Kocsor, A., Csirik, J.: A highly accurate Named Entity corpus for Hungarian. In: Proceedings of International Conference on Language Resources and Evaluation (2006)
17. Tjong Kim Sang, E. F.: Introduction to the CoNLL-2002 shared task: language-independent named entity recognition. In: Proceedings of the 6th conference on Natural language learning - Volume 20 (2002)
18. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: CONLL '03 – Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4 (2003)
19. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of HLT-NAACL 2003 (2003) 252–259
20. Zsibrita, J., Vincze, V., Farkas, R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: Tanács, A., Vincze, V. (szerk.): MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 275–283